The evolution of Ganeti, an Open Source manager for clusters of virtual machines

Michael Hanselmann

Google Switzerland

Fórum Internacional de Software Livre 11
Porto Alegre, Brazil
July 21-24, 2010
Introduction

- `getpwuid(getuid())`

Disclaimer

- Presentation content is not representative of Google’s usage of virtualization
- Presentation solely refers to the use of virtualization at Google for internal, corporate purposes and not external services or products (e.g. www.google.com)
Terminology

- Virtualization
  - [...] a hypervisor, also called virtual machine monitor, allows multiple operating systems to run concurrently on a host computer. (Wikipedia)

- Cluster

- Node ≡ physical machine

- Instance ≃ machine ≃ virtual machine
Overview

1. Introduction
2. Ganeti overview
3. Ganeti 1.x and before
4. Ganeti 2.0
5. Ganeti 2.1
6. Ganeti 2.2 and beyond
7. Additional tools
8. Questions & Answers
Ganeti overview

- Open Source cluster manager for virtualized clusters
- Combines virtualization and realtime disk replication
- Offers platform with high availability and improved resource usage
- Uses Python, OpenSSL, Xen, KVM, LVM, DRBD
- Developed at Google, opened in August 2007
- Used by Google and external users
- Licensed as GPLv2
- [http://code.google.com/p/ganeti/](http://code.google.com/p/ganeti/)
Logical Ganeti cluster organization

The evolution of Ganeti, an Open Source manager for clusters of virtual machines

Google Switzerland
Simple cluster setup

$ gnt-cluster init gntcluster1.example.net
$ gnt-node add node2
$ gnt-node add node3

$ gnt-instance add -n node1 -t plain -H memory=1G \ -s 10G -o debian web1.example.net
$ gnt-instance add -n node2 -t plain -H memory=1G \ -s 10G -o debian web2.example.net

$ gnt-instance add -n node3:node1 -t drbd \ -H memory=512 -s 20G -o redhat mail.example.net

$ gnt-instance replace-disks --auto mail.example.net
Before the beginning of time

... and out of nothing, Xencluster

- Initial experiments outside SCM
- May 26, 2006: First Perforce checkin
  - 3'124 SLOC (3'071 Python, 44 Shell, 9 Makefile)
  - Support for Xen 2.0.7
- June 21, 2006: Xencluster 1.0 released
  - 4'352 SLOC (4'220 Python, 71 Shell, 61 Makefile)
  - 41 files changed, 5'485 insertions(+), 1'217 deletions(-)
What can it do?

- Very thin layer on top of LVM and Xen command line
- Nodes hosting instances, one master node, can be failed over
- Instance disks backed directly by LVs
- Instance OS templates, only supporting instance creation
- Instance migrate $\equiv$ stop, copy data, start
- Instance backup $\equiv$ create a ready-to-run copy
But ... what is it good for?

- Designed to support office infrastructure services
- DNS, LDAP, printing, web cache, ... as virtual machines
- Drastically reduce the number of physical machines
- Keep the operational overhead of virtualization low
- Reduce the impact of hardware failure (instance backup copies)
Ganeti is [...] a wrapper around the Xen hypervisor.

— Old ChangeLog entry (May 2007)
And I name you: Ganeti!

- Real project name was needed, avoiding name collisions
- Xencluster became Ganeti
- Operational experiences & long term planning → development
- February 15, 2007: Ganeti 1.1 release
  - 6,205 SLOC (6,049 Python, 80 Shell, 76 Makefile)
  - 64 files changed, 10,922 insertions(+), 6,438 deletions(-)
Now we are getting serious

- Realtime disk replication using DRBD 0.7
- Online instance disk replacement
- Dedicated replication network for DRBD traffic (optional)
- Switch from Xen 2.0 to 3.0 series
- SMP support for instances
- Instance config changeable: memory, CPU
- Console access for instances
- `xc-watcher` to restart crashed instances
- 1.1.1 ... 1.1.9 focus on robustness improvements
And doing more with it, too!

- Converting the office infrastructure clusters
- Disk replication implies significant gain in reliability, availability
- Expanding scope: general service clusters
- Providing virtualized machines for all kinds of services
- Trial by fire or The Disk Death Incident:
  - Lots of disks kept dying for unknown reasons
  - A lot of disks to replace for several weeks
  - Turned out to be a hardware issue with the disks
  - Realtime disk replication saved the day
  - No instance data lost
Open Source, here we come!

- August 30, 2007: Ganeti goes Open Source, first beta versions of Ganeti 1.2 released
- [http://code.google.com/p/ganeti/](http://code.google.com/p/ganeti/)
- Development switched to a public Subversion repository
- Tools to support a send-patch, review by e-mail, commit cycle on top of Subversion
- December 4, 2007: Ganeti 1.2 release
  - 12'458 SLOC (12'253 Python, 122 Shell, 83 Makefile)
  - 131 files changed, 27'690 insertions(+), 12'635 deletions(-)
And we brought new toys, as well!

- More powerful OS API: create, rename, import, export
- `gnt-backup`: import/export instances
- Added DRBD 8 support, simplified replicated disk type
- Instance disk upgrade tool (DRBD 0.7 → 8)
- Cluster/node/instance tag support
- Instance reinstall & rename support
- Watcher activates disks after secondary reboot
- Scalability improvements
- Hooks: programs to be executed before/after operations
- Ganeti packages available in Debian and Gentoo
More shiny and more reliable too

- Following 1.2.x releases focus on:
  - Features: HVM support, remote API
  - Making operations easier: instance allocator, grow-disk, batched instance creation, live migration
  - Robustness: cluster-verify improvements, improved logging
  - Updates: DRBD 8.2 support
  - As well as various fixes

- Ganeti is picked up more and more by external users
- Clusters are getting larger
- November 2007: Ganeti developers start using git-svn
The only constant is change

- Ganeti 2.0: General rewrite & architectural change
- SCM changed from Subversion to Git
  (git://git.ganeti.org/ganeti.git)
- Same basic send-patch, review by e-mail, commit workflow
- May 27, 2009: Ganeti 2.0 released
  - 23,784 SLOC (23,297 Python, 313 Shell, 174 Makefile)
  - 145 files changed, 38,201 insertions(+) , 11,023 deletions(-)
Now more flexible and powerful!

- Command line tools are just a frontend now
- All work done via job queue in `ganeti-masterd`
- Master daemon as central controller, node daemons as workers
- Fine grained locking → parallelization possible
- Dropped DRBD 0.7 support, only DRBD 8 supported
- KVM supported, mixing Xen HVM & Xen PVM supported
- Read/write REST based remote API secured by basic auth & HTTPS
- Upgrade tools 1.2.7 → 2.0
Operations driving development

- Large clusters (tens of nodes) deployed internally
- 4-digit number of instances
- Cluster size provides special challenges for operations & repair
- Parallelization of Ganeti commands makes maintenance easier
- Ganeti 2.0.x:
  - Support striped LVs
  - Improved repairs
  - Robustness fixes
  - Documentation update
The latest and greatest stable release

- **Ganeti 2.1**
- March 2nd, 2010: Ganeti 2.1 release
  - 30'409 SLOC (29'223 Python, 735 Shell, 451 Makefile)
  - 145 files changed, 25’811 insertions(+), 8’630 deletions(-)
Repaired by your plastic pal

- Improved infrastructure for cluster repair (due to hardware failures)
- Infrastructure for automated disk repair
- Chroot supervisor
- Improved locking & parallelization
- More parameters for instances & hypervisors
- Documentation updates
The future is almost now

- Ganeti 2.2
  - Not there yet, first beta released on June 17, 2010
    - 41'300 SLOC (39'341 Python, 1'386 Shell, 573 Makefile)
    - 155 files changed, 25'617 insertions(+), 3'785 deletions(-)
Incoming

- Cluster merger tool
- Inter-cluster instance moves
- Inter-node RPC timeouts
- Initial support for privilege separation between daemons
- Replace SSH with SSL for instance import/export
The real future

- Development continues
- Improving cluster operations & maintenance
Wait, there is more!

htools

- Ganeti cluster allocation tools
- Started as supplementary tools for Ganeti 1.2
- Support tools for cluster operations (instance creation, repairs)
- Can talk directly to Ganeti master daemon
- Cluster rebalancer, allocator, capacity estimator
- Written in Haskell for higher performance
- git://git.ganeti.org/htools.git
The network is virtual too

- NBMA tools: Nonbroadcast Multiple Access Network tools
- git://git.ganeti.org/nbma.git
- Purpose:
  - Cluster runs in “foreign” network
  - Instances cannot be bridged to local network
  - No local IPs for instances
- Virtual network for instance traffic on top of real network
- Instance traffic is routed, not bridged
- Using GRE tunnels between nodes & gateways to outside world
Questions & Answers

Thank you for your attention.

http://code.google.com/p/ganeti/
Appendix

- Hard drive photo by Kenny Louie (cropped), http://www.flickr.com/photos/kwl/3219157599/
- SLOC (Source Lines of Code) calculated using slightly modified version of David A. Wheeler’s SLOCCount (http://www.dwheeler.com/sloccount/)
- Git (http://git-scm.com/) for diffstats
- LaTeX with Beamer (http://bitbucket.org/rivanvx/beamer/)
- Dia (http://live.gnome.org/Dia)
- gnuplot (http://www.gnuplot.info/)